



Why mobilize data?

Dmitry Schigel, Laura Russell

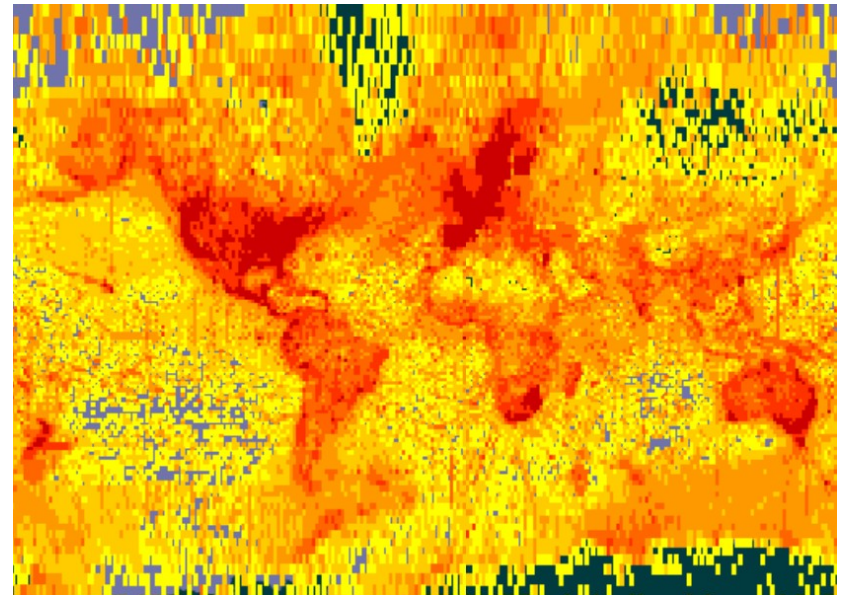
GLOBAL BIODIVERSITY INFORMATION FACILITY

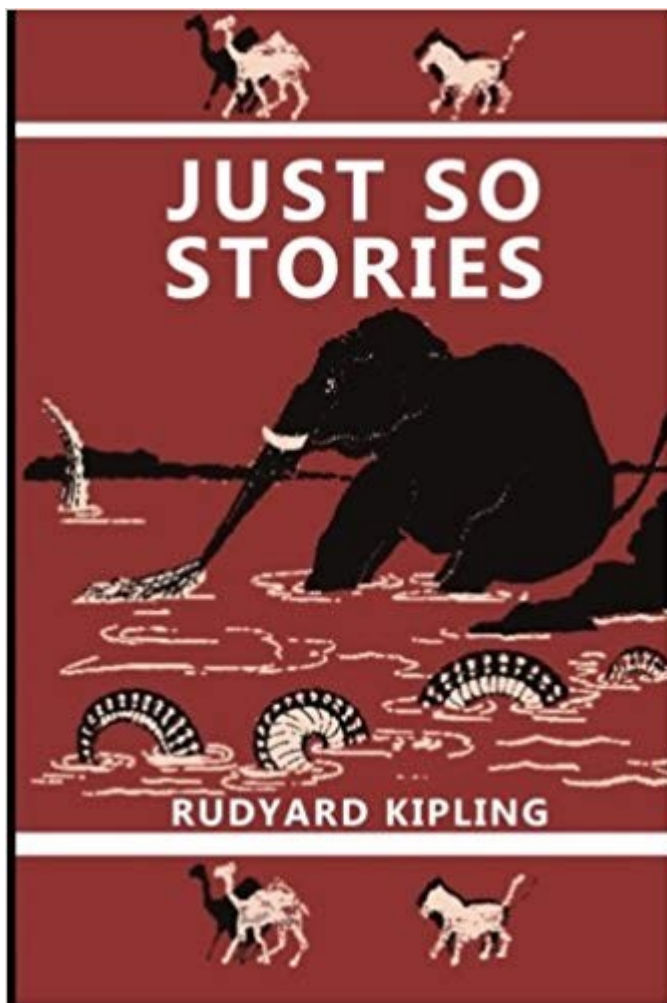


GBIF

Global Biodiversity
Information Facility

- International open data infrastructure
- Funded by the governments of the participant countries
- Network for free and open access to biodiversity data
- 92 participants:
54 countries and 39 organisations





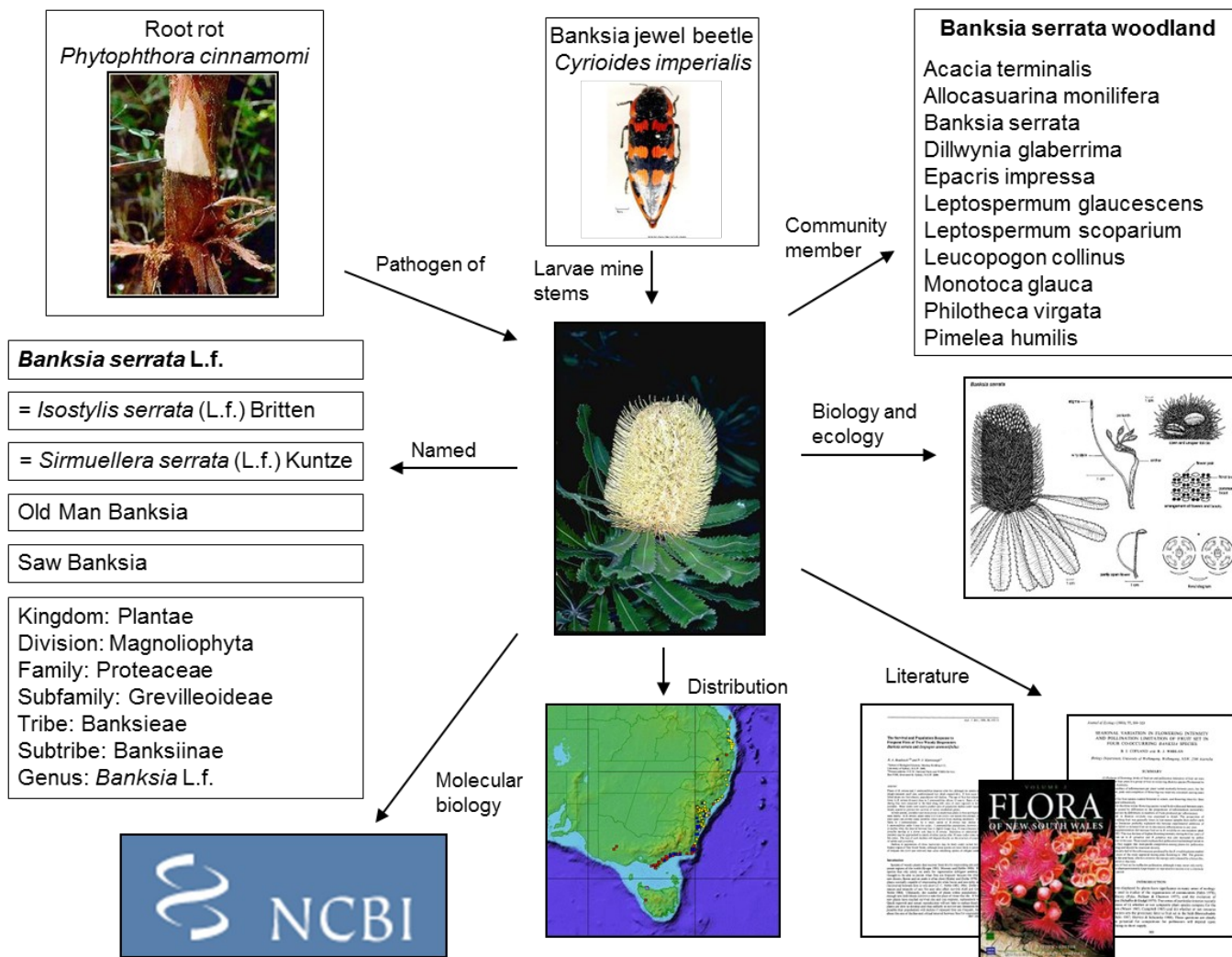
“I keep six honest serving-men
(They taught me all I knew);
Their names are **What** and **Why** and **When**
And **How** and **Where** and **Who**”

How the Elephant got his Trunk
1902

Есть у меня шестёрка слуг,
Проворных, удалых,
И всё, что вижу я вокруг,
- Всё знаю я от них.

Они по знаку моему
Являются в нужде.
Зовут их: Как и Почему,
Кто, Что, Когда и Где.

BIODIVERSITY INFORMATION



Why share data?

Why share data?

21st century = « century of the data »

Data quantity increases exponentially

Well curated and standardized, these data **have the potential** to greatly improve our knowledge and capacities

Biodiversity Data Use

Taxonomic research, niche
modelling/species distribution
prediction, invasive and alien species,
habitat degradation, interspecific
relationships, ...

But also...

Conservation biology, water
management, eco-tourism, science
history, hunting and fisheries, data
repatriation,..

Reasons to share

- Taxonomy
- Biogeographic studies
- Species diversity and populations
- Life histories and phenologies
- Endangered, Migratory and Invasive Species
- Impact of Climate Change
- Ecology, Evolution and Genetics
- Environmental Regionalisation
- Conservation Planning
- Natural Resource Management
- Agriculture, Forestry, Fisheries and Mining
- Health and Public Safety
- Bioprospecting
- Forensics
- Border Control and Wildlife Trade
- Education and Public Outreach
- Ecotourism and Recreational Activities
- Society and Politics
- Human Infrastructure Planning

GBIF ENABLED SCIENCE TOPICS

[Agriculture](#)

[Biodiversity_science](#)

[Biogeography](#)

[Citizen_science](#)

[Climate_change](#)

[Conservation](#)

[Data_management](#)

[Data_paper](#)

[Ecology](#)

[Ecosystem_services](#)

[Evolution](#)



GLOBAL RELEVANCE



**Convention on
Biological Diversity**



**GROUP ON
EARTH OBSERVATIONS**



**Biodiversity
Indicators
Partnership**



United Nations
Framework Convention on
Climate Change



Barriers to data sharing

Psychological and cultural barriers

- Lack of will
- Perceived loss of control
- Perceived data value
- Perceived data theft
- Privacy concerns

Institutional barriers

- Lack of authorization
- Lack of policies
- Business models working against data sharing

Capacity barriers

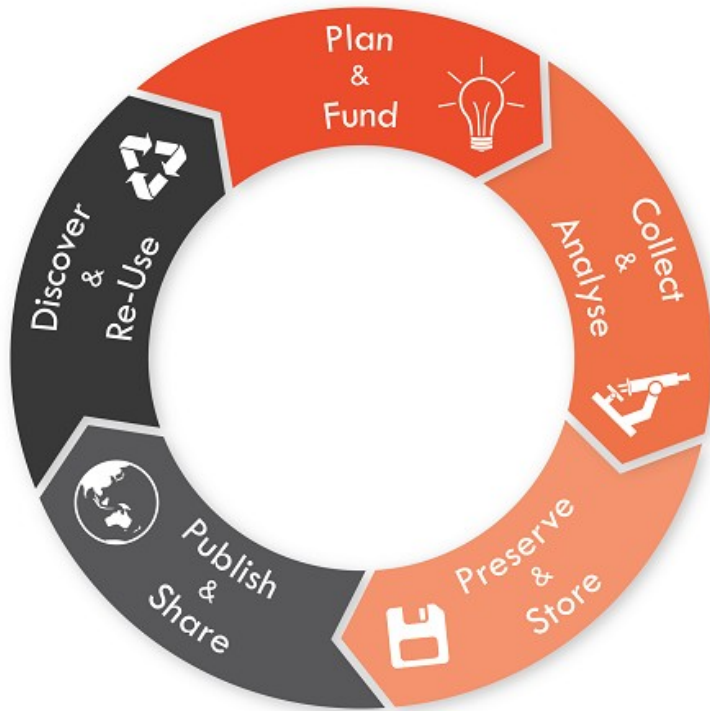
- Lack of knowledge
- Lack of understanding

Practical barriers

- Lack of funding
- Lack of infrastructure
- Lack of human resources
- Lack of time / planning

Data restriction levels

1. **Refuse to share**
2. Will only share data under **specific restrictions**
 - **Embargo**: refuse to share until they have exhausted the planned use of the data
 - **Cost**: Will only share their data for a fee
3. Agree to **share data openly**



SHORT TERM

LONG TERM

perspective



**RESEARCH
PHASE**

- file formats
- ownership
- metadata
- storage
- backups

**DISSEMINATION
PHASE**

- share with whom?
- embargo?
- licensing
- metadata

**PRESERVATION
PHASE**

- repository?
- long-term manager?

Why share data?

<https://conservationbytes.com/2018/01/07/to-share-or-not-to-share-is-no-longer-the-question/>

Corey Bradshaw

“Even if you have the intention of mining your dataset for more analyses and stacks of new manuscripts over the coming years, **making it available to the greater research community is more likely to make new opportunities** rather than stealing them away from you.”

- If you do share, at a minimum you will be cited, but you also might be invited to collaborate or co-author
- Most journals no longer allow you to be a data hoarder
- Not sharing your data can reduce your opportunities because others don't know what you've been doing

ALL SPECIMENS WILL BE DESTROYED AT SOME POINT



... but digitization and data publishing is also form of digital **security** for heritage knowledge

Get data | Share | Tools | Inside GBIF

PUBLISHER | SINCE 6 JANUARY 2014

Museu Nacional / UFRJ

ABOUT | HOME PAGE

26,660 OCCURRENCES | 28 DATASETS | 44 CITATIONS

Description: The Museum shelters one of the largest exhibits of the Americas, consisting of animals, insects, minerals, aboriginal collections of utensils, Egyptians mummies and South American archaeological artifacts, meteorites, fossils and many other findings.

Endorsed by: GBIF Brazil

Installations: IPT do Museu Nacional do Rio de Janeiro / UFRJ

Administrative contact: Claudia Rodrigues Ferreira de Carvalho

Technical contact: Cristiana Silveira Serejo

Country or area: Brazil

Hosting: 17 datasets (1 publisher • 1 country)

129 OCCURRENCES WITH IMAGES

Museu Nacional / UFRJ
Brazil
<http://www.museunacional.ufrj.br/>

Cristiana Silveira Serejo
Technical point of contact
cristiana@gbif.org.br

Claudia Rodrigues Ferreira de Carvalho
Administrative point of contact
claudia@mn.ufrj.br

52,584 GEOREFERENCED RECORDS

Generated 4 hours ago © OpenStreetMap contributors © OpenMapTiles © OpenStreetMap contributors

Any year 1819 - 2016 EXPLORE

What is GBIF? | API | FAQ | Newsletter | Privacy | Terms and agreements | Citation | Acknowledgements

Contact | GBIF Secretariat Universitetsparken 15 | DK-2100 Copenhagen Ø | Denmark

GBIF Global Biodiversity Information Facility

f t in v

How to share your data?

DISCOVERY

Bibliographic ->

Include **tables** into your publication
Table 1, Table 2, Supplementary materials

General ->

Publish tables on a **standalone website**
Institutional, thematic, personal pages

General ->

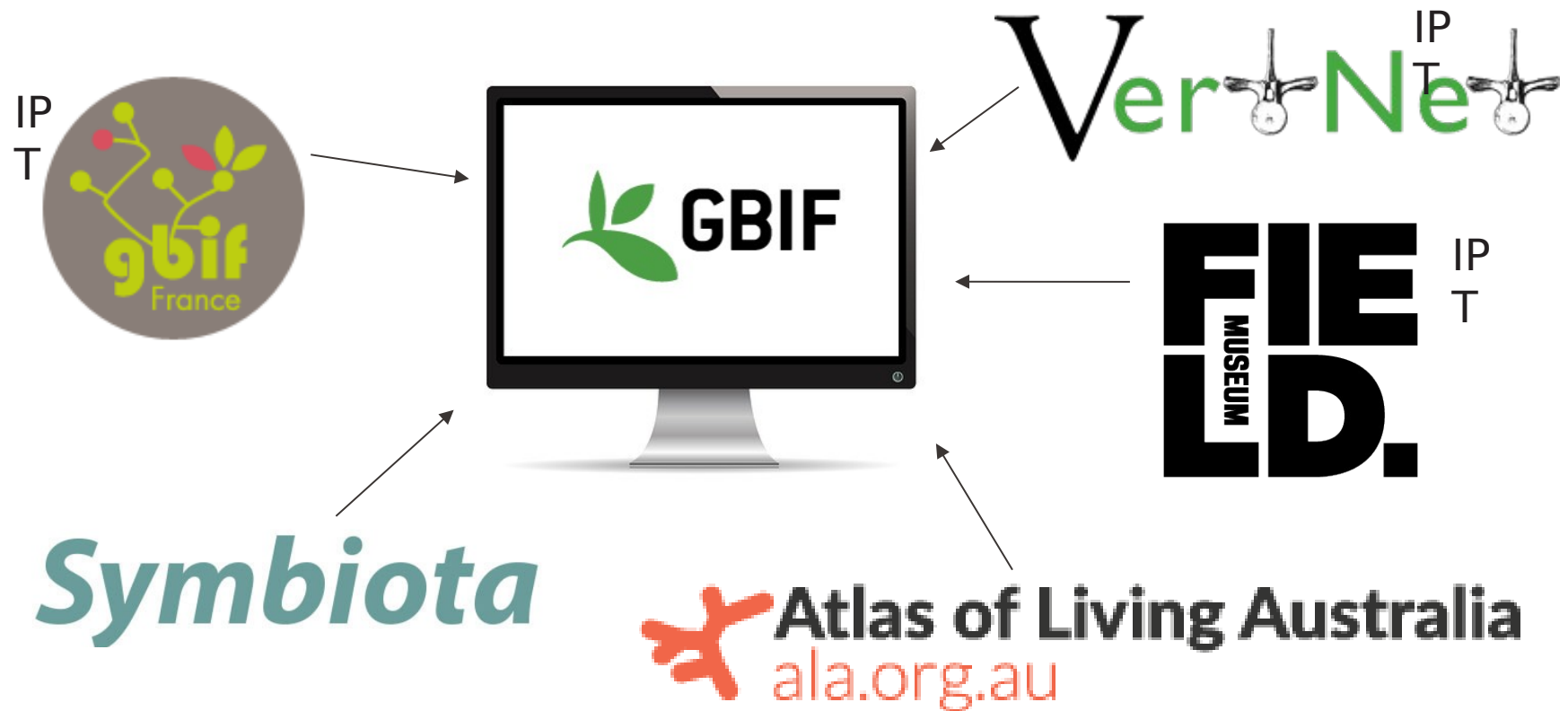
Archive as-is in a data repository
Dryad, FigShare, Zenodo etc.

Specific, filtered,
->
API access

Format your data to comply with
international data standards and
publish through global portals
NCBI Genbank, GBIF, OBIS


What is Data Publishing?

“Data Publishing” refers to making biodiversity datasets publicly accessible and discoverable, in a standardized form, via an access point, typically a web address (a URL).”













What Does IPT Stand For?

Integrated Publishing Toolkit


GBIF INTEGRATED PUBLISHING TOOLKIT (IPT)
free and open access to biodiversity data

Hosted resources available through this IPT

Filter:

Logo	Name	Organisation	Type	Subtype	Records	Last modified	Last publication	Next publication
	Field Museum of Natural History (Botany) Seed Plant Collection	Field Museum of Natural History	Occurrence	Specimen	576,367	2018-03-05	2018-03-05	--
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	115,932	2018-03-05	2016-12-19	--
	Field Museum of Natural History (Botany) Fungi Collection	Field Museum of Natural History	Occurrence	Specimen	62,851	2018-03-05	2018-03-05	--
	Field Museum of Natural History (Botany) Lichen Collection	Field Museum of Natural History	Occurrence	Specimen	55,370	2018-03-05	2018-03-05	--
	Field Museum of Natural History (Botany) Pteridophyte Collection	Field Museum of Natural History	Occurrence	Specimen	70,784	2018-03-05	2018-03-05	--
	Field Museum of Natural History (Geology) Fossil Invertebrates Collection	Field Museum of Natural History	Occurrence	Specimen	62,149	2018-01-23	2017-01-06	--
	Field Museum of Natural History (Geology) Paleobotany Collection	Field Museum of Natural History	Occurrence	Specimen	22,851	2018-01-23	2017-01-27	--
	Field Museum of Natural History (Zoology) Amphibian and Reptile Collection	Field Museum of Natural History	Occurrence	Specimen	285,342	2018-01-23	2017-02-06	--
	Field Museum of Natural History (Zoology) Bird Collection	Field Museum of Natural History	Occurrence	Specimen	527,634	2018-02-27	2017-05-04	--
	Field Museum of Natural History (Zoology) Bird Egg Collection	Field Museum of Natural History	Occurrence	Specimen	20,992	2018-02-27	2018-02-27	--

Data -> IPT -> GBIF

GBIF INTEGRATED PUBLISHING TOOLKIT (IPT)
free and open access to biodiversity data

Home About

Hosted resources available through this IPT

Logo	Name	Organisation	Type	Subtype	Records	Last modified	Last publication	Next publication
	Field Museum of Natural History (Zooology) Insect, Arachnid and Myriapod Collection	Field Museum of Natural History	Occurrence	Specimen	576,367	2018-03-05	2018-03-05	—
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	62,851	2018-03-05	2018-03-05	—
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	55,370	2018-03-05	2018-03-05	—
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	70,784	2018-03-05	2018-03-05	—
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	399,688	2018-03-05	2018-03-05	—
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	252,069	2018-03-05	2018-03-05	—
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	20,992	2018-02-27	2018-02-27	—
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	216,325	2018-01-26	2018-01-26	—
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	109,845	2017-11-15	2017-11-15	—
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	527,634	2018-02-27	2017-05-04	—
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	285,342	2018-01-23	2017-02-06	—
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	22,851	2018-01-23	2017-01-27	—
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	62,149	2018-01-23	2017-01-06	—
	Field Museum of Natural History (Botany) Bryophyte Collection	Field Museum of Natural History	Occurrence	Specimen	115,932	2018-03-05	2016-12-19	—

Showing 1 to 14 of 14

The most recently updated resources are also available as an [RSS feed](#)

IPT Version 2.3.5-rb9b0544 [About the IPT](#) [User manual](#) [Report a bug](#) [Request new feature](#)

©2017 Global Biodiversity Information Facility

Field Museum of Natural History (Zooology) Insect, Arachnid and Myriapod Collection
Published by Field Museum
Sharon Grant • Crystal Maier

999,688 Occurrences 11 Citations

90% With taxon match 23% With coordinates 72% With year

World map showing distribution of specimens (yellow dots).

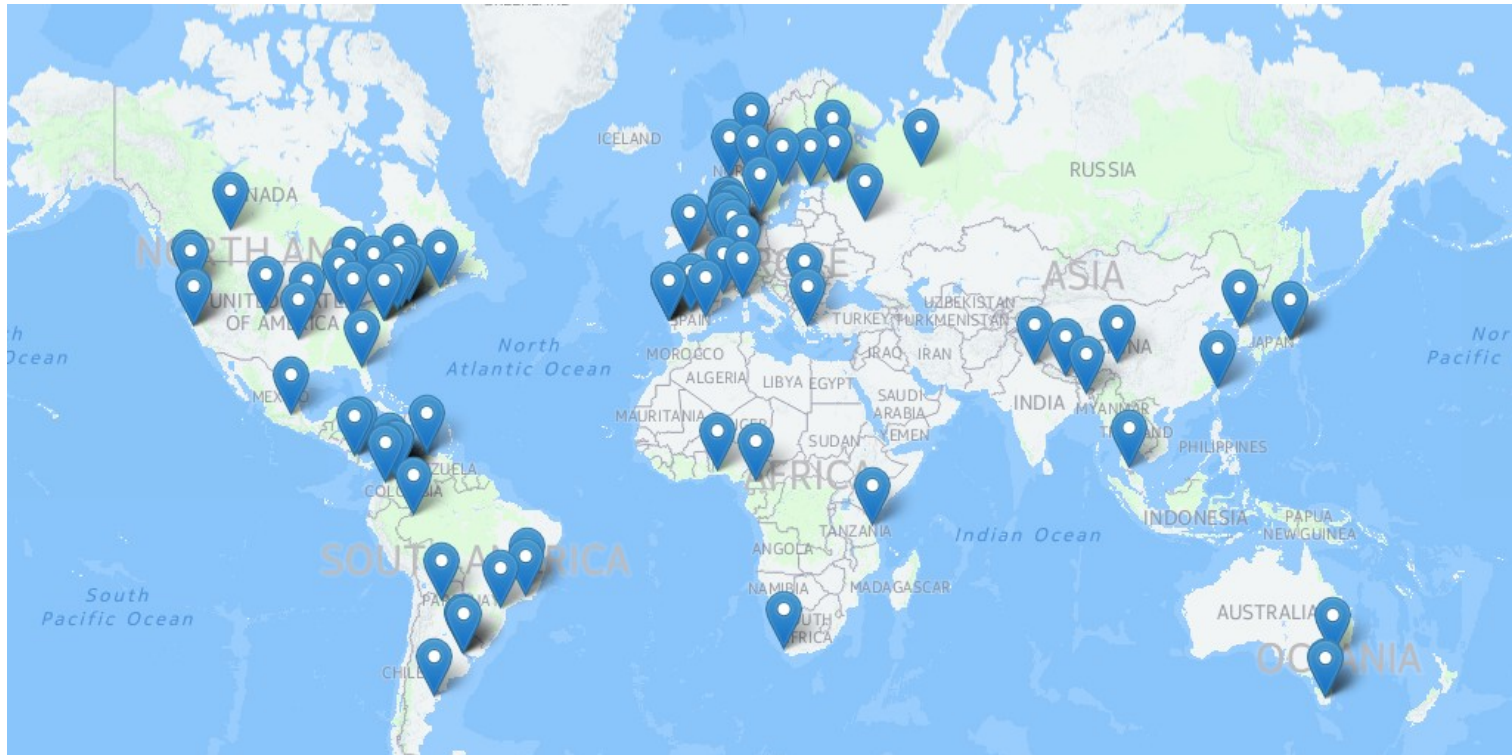
Description
The Division of Insects' holdings of worldwide Arthropoda (excluding Crustacea) rank fifth in overall size among North American collections and are of worldwide importance for many groups. The collection presently includes roughly 4.1 million pinned insects plus 6.3 million specimens or lots in alcohol or on microscope slides. In addition, there are over 17,000 partly-sorted "bulk samples" from traps or leaf-litter extractions. The collection receives heavy use by US and international research visitors and borrowers as well as extensive educational use.

Geographic coverages
Global

Additional info
<https://www.fieldmuseum.org/field-museum-natural-history-conditions-and-suggested-norms-use-collections-data-and-images>

Contributors
Crystal Maier
Metadata Author
Collection Manager Insects, Arachnids and Myriapods
Field Museum of Natural History
1400 S Lake Shore Drive
Chicago 60605
United States
crystal@fieldmuseum.org

IPT Stats: *Dec 2016*



174 installations

52 countries

211 checklists

2,851 occurrence datasets

55 sampling event datasets

100+ million records

BENEFITS OF OPENESS

Increases the efficiency of research

Promotes scholarly rigor and quality of research

Enables tracking of data use and data citation through DOIs

Expands the spectrum of academic products through data papers

Enhances visibility and scope for engagement

Enables researchers to ask new research questions

Enhances collaboration and community-building

Increases the economic and social impact of research

International conventions and requirements from funding agencies

Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar*, Roger S. Day, Douglas B. Fridsma

Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America

Background. Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available. **Principal Findings.** We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly ($p=0.006$) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression. **Significance.** This correlation between publicly available data and increased literature impact may further motivate investigators to share their detailed research data.

Citation: Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

INTRODUCTION

Sharing information facilitates science. Publicly sharing detailed research data—sample attributes, clinical factors, patient outcomes, DNA sequences, raw mRNA microarray measurements—with other researchers allows these valuable resources to contribute far beyond their original analysis[1]. In addition to being used to confirm original results, raw data can be used to explore related or new hypotheses, particularly when combined with other publicly available data sets. Real data is indispensable when investigating and developing study methods, analysis techniques, and software implementations. The larger scientific community also benefits: sharing data encourages multiple perspectives, helps to identify errors, discourages fraud, is useful for training new researchers, and increases efficient use of funding and patient population resources by avoiding duplicate data collection.

Believing that that these benefits outweigh the costs of sharing research data, many initiatives actively encourage investigators to make their data available. Some journals, including the *PLoS* family, require the submission of detailed biomedical data to publicly available databases as a condition of publication[2–4]. Since 2003, the NIH has required a data sharing plan for all large funding grants. The growing open-access publishing movement will perhaps increase peer pressure to share data.

However, while the general research community benefits from shared data, much of the burden for sharing the data falls to the study investigator. Are there benefits for the investigators themselves?

A currency of value to many investigators is the number of times their publications are cited. Although limited as a proxy for the scientific contribution of a paper[5], citation counts are often used in research funding and promotion decisions and have even been assigned a salary-increase dollar value[6]. Boosting citation rate is

RESULTS

We studied the citations of 85 cancer microarray clinical trials published between January 1999 and April 2003, as identified in a systematic review by Ntzani and Ioannidis[7] and listed in Supplementary Text S1. We found 41 of the 85 clinical trials (48%) made their microarray data publicly available on the internet. Most data sets were located on lab websites (28), with a few found on publisher websites (4), or within public databases (6 in the Stanford Microarray Database (SMD)[8], 6 in Gene Expression Omnibus (GEO)[9], 2 in ArrayExpress[10], 2 in the NCI GeneExpression Data Portal (GEDP)(gedp.nci.nih.gov); some datasets in more than one location). The internet locations of the datasets are listed in Supplementary Text S2. The majority of datasets were made available concurrently with the trial publication, as illustrated within the WayBackMachine internet archives (www.archive.org/web/web.php) for 25 of the datasets and mention of supplementary data within the trial publication itself for 10 of the remaining 16 datasets. As seen in Table 1, trials published in high impact journals, prior to 2001, or with US authors were more likely to share their data.

The cohort of 85 trials was cited an aggregate of 6239 times in 2004–2005 by 3133 distinct articles (median of 1.0 cohort citation per article, range 1–23). The 48% of trials which shared their data received a total of 5334 citations (85% of aggregate), distributed as shown in Figure 1.

Academic Editor: John Ioannidis, University of Ioannina School of Medicine, Greece

Received: December 13, 2006; **Accepted:** February 26, 2007; **Published:** March 21, 2007

Piwowar et al.
(2007)
Content CC-BY-2.0

DATA ACCESS – FREE AND OPEN TO ALL

GBIF.org: search, browse a download

- Occurrences
- Species
- Datasets
- Publishers
- Countries

GBIF API – machine friendly e-access

- External systems

The screenshot displays two pages from the GBIF.org website. The top page is titled "Citation guidelines" and dated "14 OCTOBER 2016". It explains the purpose of citation guidelines and provides examples of how to cite GBIF data. The bottom page is titled "API Summary" and dated "DEVELOPER | API DOCS". It provides a summary of the GBIF API, including its base URL and a list of API sections.

Citation guidelines
 14 OCTOBER 2016
 These guidelines provide the most common examples of citation by GBIF users.

The practice of citation serves two primary purposes: to acknowledge the original source of information and to help other researchers find that source. As an open data research infrastructure, GBIF encourages good citation practices to ensure proper credit and attribution as well as transparency and reproducibility.

Below you'll find guidelines for the most common cases of citation by GBIF users. While these are presented in Harvard style, please feel free to adapt citations to the style format required by your institution, publisher or agency. However, please do include each element of content from the relevant example, especially the **DOI link, URL, and date**.

Citation examples

Occurrence data download through GBIF.org

When a registered user downloads data from GBIF.org, s/he is redirected to a page that includes the following citation:

When using this dataset please use the following citation:
 GBIF.org (29th February 2016) GBIF Occurrence Download <http://doi.org/10.15468/8t-ywqpmv>

This citation also appears in a confirmation sent to the email account that the user registered with.

By using the assigned DOI included with your citations, you vastly improve GBIF's ability to track the use of data, which we can then report to data publishers. It also provides the mechanism for connecting published uses of the data back to its sources. In addition to acknowledging them, the practice of using DOI citations rewards publishers by reinforcing the value of sharing open data to the publisher's stakeholders and funders.

Individual checklist, occurrence or sampling event dataset

Each dataset page contains a paragraph that provides a default citation, for example:

DEVELOPER | API DOCS
API Summary
<http://api.gbif.org/v1/>

SUMMARY **REGISTRY** **SPECIES** **OCCURRENCE** **MAPS** **NEWS**

The GBIF API is a RESTful JSON based API. The base URL for v1 you should use is <http://api.gbif.org/v1/>. The API should be considered stable, as should this accompanying documentation. Please report any issues you find with either the API itself or the documentation using the "Feedback" button on the top right.

Content
 Sections
 Communication
 Common operations
 Authentication
 Enumerations
 Roadmap to v2

API Sections
 The API is split into logical sections to ease understanding.
Registry: Provides means to create, edit, update and search for information about the datasets, organizations (e.g. data publishing networks and the means to access them (technical endpoints)). The registered content controls what is crawled and indexed in the GBIF data portal, but as a shared API may also be used for other initiatives.
Species: Provides services to discover and access information about species and higher taxa, and utility services for interpreting names and looking up the identifiers and complete scientific names used for species in the GBIF portal.
Occurrence: Provides access to occurrence information crawled and indexed by GBIF and search services to do real time paged search and asynchronous download services to do large batch downloads.
Maps: Provides sample services to show the maps of GBIF mobilized content on other sites.
Workflow/Tools: Provides services to automate content publishing and access to GBIF content via GBIF workflow execution API.

Communication
 You can sign up to the GBIF API users mailing list to post your questions, and to keep informed about the API. We will announce new versions, and scheduled maintenance downtimes before they happen.
 We welcome any example uses of the API to guest feature on the GBIF developer blog.
 Feedback from developers on the API can be sent to informatics@gbif.org.

Common operations
 The following details common cross-cutting parameters used in the API.

Paging
 For requests that support paging the following parameters are used:

How does it work (for you)?

Seven steps of publishing data through GBIF:



1. **Agree** with your administration
2. **Register** your institution / collection
3. **Understand** what data type you are dealing with (4 types)
4. **Standardize** / format your data (export)
5. **Choose** your way to publish
6. **Check** your data
7. **Publish!**

GBIF member nodes endorse institutions to share data

- Optionally enforcing data quality thresholds
- Judgment on a case by case basis

Institutions publish datasets on GBIF

- Retain ownership
- Referenced as the “publisher”

Institutions often offer assistance to others in separate agreements

- Hosting services

Choose your data license

- CC0 and CC BY - 90% datasets, 85% records
- CC BY-NC - 7% of datasets, 10% of records

THANK YOU



**Global Biodiversity Information
Facility (GBIF)**

Secretariat
Universitetsparken 15
DK-2100 Copenhagen Ø
DENMARK



Discussion

Discussion

What are the barriers to open sharing of data from your areas?

How do you think these barriers can be overcome?

